Commodity tags 1.0

The aim is to provide labels to a sample of scrapped articles and to find their common features using a term frequency approach.

Data

60934 scrapped articles compose initial data: they were scrapped on 28/07/2016. A 4000 observations random sample is taken out from the population and is analyzed.

Four lists of given stemmed words are used in the classification task. The lists are:

Class	Label	Words List
Class 0	Wheat	Wheat, crop, grain, cereal, grass, durum, gluten, semolina, spelt, harvest, farm, agricoltur
Class 1	Maize	maiz, corn, crop, grain, cereal, harvest, farm, agricoltur
Class 2	Soybeans	soybean, soy, crop, harvest, farm, agricoltur
Class 3	Rice	rice, paddi, crop, grain, harvest, farm, agricoltur, mill, kernel

These words have been sorted out from the most common (stemmed) words used to describe wheat, maize, rice and soybeans.

Procedure

The script is written in Python.

First of all the articles are loaded from the jsonl file and a 4000 articles sample is randomly sorted out. The sample is then tokenized, cleaned from its stop words and stemmed using the Snowball stemmer.

Unfortunately it may happen that in the sample, an empty article in the form of an empty list is present. Since this occurrence stops the script during the

term frequencies count, empty articles will be replaced by the NA value represented by the string:

```
['NA','NA','NA','NA','NA','this','this','this','this','article','article','article'
,'is','is','missing']
```

By doing so, the resulting empty article's five most used words are going to be: "Na", "this", "article", "is" and "missing".

After article tokenization, stop words removal, stemming and NA identification, articles are then classified by counting the number of times certain words appear: these words are the initial words given in the "Data" section of this paper. The lists are applied to stemmed articles in order to limit the set of possible words, namely to avoid missing the word "soybeans" for including in the initial lists just the word "soybean".

Article's possible classes are:

Class	**Label**
Class 0	Wheat
Class 1	Maize
Class 2	Soybeans
Class 3	Maize
Class 4	General Agriculture
Class 5	Other

An article is then classified in one of the above classes depending on their words counters: "wheat" (class 0), "maiz" (class 1), "rice" (class 2) and "soybean" (class 3) strongly define the class helped by related words, as "paddi" or "kernel" for rice, while general words as "agriculture" are included in all the lists in order to discriminate agriculture related articles from non-agriculture related ones. When no initial lists word is found or the sum of all the critical words is below a significancy threshold (defined as 10), the article is classified as class 5. Once done that, the ratio of agriculture related articles over the total and the contained ratio of wheat, maize, soybeans and rice related articles over the total are computed. The ratios are respectively named Commodity Related Articles (CRA) and Agriculture Related Articles (ARA).

Once computed the ratios, the five most used words for each article are provided: column "Word1" gives the most used word, "Word2"returns the second most used word and so on. Most used words are extracted in order to

understand which other words may be used as features, but they can also be used to understand each article context or can be thought as additional labels.

At the end of the script all this information is resumed in a .csv file called *df.csv:* in the file the dataframe shows as columns “article number”, “article index”, “wheat” word counter, “maiz” word counter, “soybeans” word counter, “rice” word counter, assigned class, word1, word2, word3, word4 and word5.

Results

On a 4000 observations sample, this analysis resulted in 24.4% of the articles labeled as agriculture related, which 22.6% are wheat, maize, soybeans and rice related. In terms of the ratios:

Ratio	**Percentage**
CRA	24.4%
ARA	22.6%

The remaining part is classified as general. Given the number of observations contained in the sample, running the script again will result in similar ratios.

Most of the articles are classified as class 0 or class 5, as shown in the table.

Class	**Label**	**Articles**
Class 0	Wheat	798
Class 1	Maize	87
Class 2	Soybeans	11
Class 3	Rice	9
Class 4	General Agriculture	69
Class 5	Other	3023

What it appears from a first inspection is that many of class 0 and 1 articles have high “wheat” and “maize” counters and they also show little counters differences: this means that class 0 articles show an high number of agriculture related words (high counters) and many articles talk about both wheat and maize (little difference) as other commodities. To support this thesis can be said that a significant proportion of class 0 articles has as Word1 “wheat” and

as Word2 or Word3 “corn”. Given this evidence a multiple label approach is advised.

Moreover it seems that many class 0, 1 articles have to be properly formatted, because they are quite raw. An example of this kind of articles can be the following:

home commodities companies markets opinion data calendar Thurs 28 Jul 2016 15:34 UK, 16 Jul 2014, by Agrimoney.com Fears over China sorghum import curbs \'overplayed\' Fears that China may restrict its soaring sorghum imports,\rafter already clamping down on trade in corn and distillers\' grains, may be\rmisplaced, US farm officials said, explaining an upgrade to its forecast for buy-ins. The US Department of Agriculture noted industry reports\rsuggesting that China might increase regulation of its booming sorghum imports,\rwhich are expected to rise fivefold to a record 3.40m tonnes in 2013-14. China has already clamped down on buy-ins of corn and distillers\'\rgrains (DDGs), a corn-derived feed ingredient, ostensibly over concerns of\rcontamination with a genetically modified variety approved in Washington and Beijing,\ralthough many investors suspect that the move is aimed at encouraging use of\rthe country\'s own large supplies. After a succession of strong harvests, China\'s corn stocks\rwill hit 87.3m tonnes, equivalent to nearly half the world total, at the end of\r2014-15, the International Grains Council believes. \'Not substantiated\' However, ideas of curbs on Chinese sorghum imports "have not\ryet been substantiated", USDA officials said. Furthermore, "the prices of interior or export positions do\rnot appear to reflect a resultant policy change". In fact, the officials noted that Chinese buyers have been bidding\rup for sorghum because it faces less regulatory red tape, in part because it is\rgrown from conventional, rather than genetically engineered, seed - unlike\rcorn. "As a non-GM crop, it faces less barriers getting through customs\rin China, and no import quota is required." Export competition The comments came as the USDA expanded on a 200,000-tonne\rupgrade to 3.9m tonnes in its forecast for Chinese sorghum imports in 2014-15. That would represent a 500,000-tonne rise on this season\'s\rrecord volumes, and represent four times as much as China has imported over the\rprevious 50 year combined. The demand is helping support sorghum exports from the US,\rwhich are expected at a 10-year high of 5.0m tonnes in 2013-14, falling only\rmodestly to 4.5m tonnes next season, 500,000 tonnes more than previously\rthought. Conversely, the USDA has cut by 200,000 tonnes to 800,000\rtonnes its forecast for Australian sorghum shipments in 2014-15, reflecting the\rimpact of drought on its harvest, "which is constraining exports". \r/* customize styles to your web-site */\r.mp_releases_box {\r border: solid #999999 1px;\r background-color: #FFFFFF;\r}\r#mp_releases {\r font-family:Geneva, Arial, Helvetica, sans-serif;\r padding-top: 0px;\r position: relative;\r vertical-align: bottom;\r}\r#mp_releases li {\r margin-top: 3px;\r}\r#mp_releases a {\r color: #0033CC;\r font-size: 12px;\r text-decoration: underline;\r}\r#mp_releases a:hover {\r color: #CC0000;\r font-size: 12px;\r}\r.ferthead {\r font-size: 14px;\r color: #000000;\r}\r.fertheadbox {\r padding-left: 10px;\r padding-top: 10px;\r}\r Ideas wane on Argentine wheat revival - but corn hopes grow Hedge funds sell ags again - but wheat shorts \'in trouble\' Analysts warn of cotton market \'correction\' as Chinese imports slump Brexit threat to ag prices not over yet Dryness takes further toll on Brazil\'s robusta crop Starbucks slows pace of coffee buying, even as prices rise PM markets: funds pile out of soybean longs, wheat shorts CBH plans \'emergency\' storage in face of bumper Aussie grains harvest Corn Belt farmland prices fall again, despite crop price firmness \r \r 10. European wheat yield hopes lifted, but not in France \'Era of high ag prices quite likely over\' - OECD, UN Grain prices rebound, after US data on corn, wheat stocks fall short Hedge funds accelerate ag selldown - boding well for wheat prices? Hedge funds quids in, as favouring softs over grains comes good La Nina \'to boost\' prospects for South American wheat harvest Corn, soybean futures rally on \'searing\' forecast for US Midwest Corn futures tumble, wheat hits 9-year low, as US lifts supply hopes UN

cuts corn output forecast - as wheat prospects improve Corn, soybean futures tumble anew as La Nina fears subside \r \r 10. Hedge funds\' sell-off in grains raises hopes for corn, wheat rallies Wheat, coffee, cotton futures manage firm end to poor month Hedge funds turn record bearish on hard wheat - raising concerns Hedge funds cut bullish ag bets, amid China, meat cancer fears CF backs ideas of rise in US corn sowings in 2016 Food prices rise at fastest in three years, says UN SocGen upbeat on wheat, soybean, coffee prices US farmers one-third slower selling corn than normal - ADM US ethanol output soars, defying \'very weak margin\' talk Wheat futures - will they gain in 2016 for the first time in four years? \r \r 10. Hedge funds \'may have overreacted\' in record wheat sell-down Cocoa futures jump despite sharp drop in Asia\'s grind Wheat soars, helping corn higher. But coffee, cotton drop Cocoa futures rise on better-than-expected European demand Cattle futures ease, as feedlot numbers beat forecasts European dairy futures could finally catch on ABN downbeat on cocoa, coffee prices, citing demand doubts Wheat, coffee, cotton futures manage firm end to poor month Evening markets: corn, cotton futures drop as Wasde looms Wheat prices - will they continue their collapse in 2014? \r \r 10. PM markets: wheat futures dodge bullet, but other ags drop Home About RSS Commodities Companies Markets Subscribe Legal disclaimer Privacy policy Contact Acknowledgements \r//Start Tab Content script for UL with id="maintab" Separate multiple ids each with a comma.\rinitializetabcontent("wholetab");\r \r var _bizo_data_partner_id = "513";\r \r var _bizo_p = (("https:" == document.location.protocol) ? "https://sjs." : "http://js.");\r document.write(unescape("%3Cscript src=\'" + _bizo_p + "bizographics.com/convert_data.js?partner_id=" + _bizo_data_partner_id + "\' type=\'text/javascript\'%3E%3C/script%3E"));\r

Inspecting most used words, it has to be underlined that many class 0 and 1 articles are also related to “trade” and “futures”.

On the other hand class 5 articles are quite noisy, in the sense that in class 5 can be found an article that deals with the Vatican bank:

The Vatican bank disclosed its annual report for the first time in the institutes history as it seeks to improve financial transparency after several corruption scandals. The bank, formally called the Institute for Works of Religion, or IOR, expects 2013 to be marked by extraordinary expenses for the ongoing reform and remediation processand the effects of rising interest rates, according to a statement. The bank earlier this year reported that 2012 profit more than quadrupled to 86.6 million euros ($117 million). A review of all customer relationships and procedures to prevent money-laundering is under way, the bank said in the statement posted on its IOR will shut down about 900 accounts, including all of those held by foreign embassies, Corriere della Sera reported today without saying how it got the information. The decision was taken after viewing large cash transactions by diplomatic missions of Iran, Iraq and Indonesia, according to the newspaper. The remediation efforts and the introduction of appropriate regulation in the institute apply independently of the nature of the clients, Max Hohenberg, a spokesman for the IOR, said by phone. It doesnt matter whether they are employees, cardinals or ambassadors. Todays annual report publication is part of Vatican bank President Ernst von Freybergs effort to help transform a 71-year-old institution rocked by scandal into a transparent financial firm. In June, Pope Francis named a special commission to help oversee the operations of the bank after Moneyval, the Council of Europes monitoring body for money laundering and terrorism financing, called for independent supervision of the bank. Von Freyberg assumed responsibility for 18,900 clients when he accepted the appointment from Franciss predecessor, Pope Benedict XVI, in February. His team is reviewing the source of all deposits, from the savings of individual nuns and priests to the operating resources of Catholic congregations with worldwide reach. The bank oversees about 7.1 billion euros in assets, largely in bonds and cash. IOR clients with outposts from Chile to Tanzania manage income, transfers and

expenditures out of the bank located in the worlds smallest state, situated in the heart of Rome. The bank doesnt use deposits for lending and had less than 1 billion euros in equities at the end of last year. The banks profit is at the disposal of the Holy See. Last year it gave the Pope a contribution of 50 million euros. The IOR had 114 employees at the end of 2012 and is housed in a building adjacent to the papal offices just off St. Peters Square. The Vatican is trying to overcome three decades of scandals ranging from the Banco Ambrosiano failure in the 1980s to the freezing of 23 million euros by Italian prosecutors in 2010 in a money-laundering probe. While admitting no wrongdoing, the Vatican paid $240 million to Banco Ambrosiano account holders in 1984 after the IOR was implicated in the lenders fraudulent bankruptcy. Ambrosianos former chairman Roberto Calvi, dubbed Gods banker, was found hanged under Londons Blackfriars Bridge in June of 1982 amid the scandal. The Vatican signed an agreement with Italian authorities July 26 to exchange information about the bank in order to prevent money laundering and terrorism financing. Similar accords have also been signed with other countries, including the U.S.'

an article related to russian oil production

Lukoil PJSC, Russias second-largest oil producer, said first-quarter profit dropped 59 percent as crude prices declined to a 12-year low. Net income fell to 42.8 billion rubles ($651 million) from 104 billion rubles a year earlier, the Moscow-based company said in a statement on Monday. That beat the 41.3 billion-ruble estimate of six analysts surveyed by Bloomberg. Free cash flow declined 43 percent to 36 billion rubles. Free cash flow is the weakest point in the entire first-quarter report, said Alexander Kornilov, an oil and gas analyst at Aton. Free cash flow is the key thing investors watch in Lukoil numbers, making judgments about the companys capability to increase its dividend payments in ruble terms. Russian producers have been partially buffered against the rout in crude by a weaker ruble, which has reduced costs, and taxes that decline with lower prices. While higher output helped smaller rivals Bashneft PJSC and Gazprom Neft PJSC boost profit in the first quarter, Lukoils production in Russia has dropped. Lukoil rose 1 percent to 2,613 rubles as of 11:31 a.m. in Moscow trading as Brent crude climbed to more than $50 a barrel. Rosneft OJSC, Russias largest producer, advanced 1.7 percent, while third-ranked OAO Surgutneftegas gained 1.8 percent. Oil and gas output fell to 2.35 million barrels of oil and gas a day, Lukoil said. Production at the companys largest Western Siberia unit dropped 8.4 percent to 865,000 barrels a day as the fields age and Lukoil invests in higher return projects, it said. Output increased in Iraq, where Lukoil has so far received$5.6 billion of oil in compensation for project costs of $7.07 billion, according to the statement. The company was also paid a further $231 million in Iraq. Lukoil spent 4.5 billion rubles on exploration and production in Nigeria, where the company could eventually pump 6 million to 7 million tons of oil a year, Chief Executive Officer Vagit Alekperov told Russian state television station Rossiya-24. Earnings before interest, taxes, depreciation and amortization fell to 145 billion rubles, missing the 158 billion-ruble estimate of seven analysts. Revenue fell to 1.18 trillion rubles, Lukoil said.

or the status of Palestine – Israel international relations

Israeli Prime Minister Benjamin Netanyahu canceled a visit to Germany as violence between Palestinians and Israelis continued despite steps to curb it. A Palestinian man was shot dead in the southern Israeli city of Kiryat Gat Wednesday after he stabbed a soldier and tried to flee with his rifle, police said. Earlier in the day, a Palestinian woman was shot and seriously wounded after she tried to stab a Jewish passerby in Jerusalems old city. The series of attacks prompted Netanyahu to cancel a meeting in Berlin with German Chancellor Angela Merkel scheduled for Thursday, his office said by e-mail. The prime minister will remain in Israel to closely monitor the situation, according to the statement. Police reported a third incident later on Wednesday in which a Palestinian stabbed an Israeli man outside a mall in the Tel Aviv suburb of Petach Tikva. The victim suffered moderate injuries,

police spokeswoman Luba Samri said in a text message. Israeli and Palestinian leaders have blamed each other for the upsurge in tension, much of it stemming from disagreement over a Jerusalem shrine, known to Jews as the Temple Mount and as the Noble Sanctuary to Muslims. Four Israelis and four Palestinians have been killed over the past week, and dozens more have been wounded on both sides in incidents and clashes in Jerusalem and the West Bank. Netanyahu has ordered new security measures in a bid to stop the violence, including easing open-fire orders for Israeli troops and expediting the demolition of homes of Palestinians convicted of attacks. Palestinian President Mahmoud Abbas said Wednesday that Israel is to blame for the unrest, citing Jewish settlement growth, attacks on Palestinian property by Jewish extremists, and what he called Israeli aggression at the holy site. Israeli President Reuven Rivlin told a press conference Thursday that some radical Muslim elements, including an Islamic movement based in the countrys north, are deliberately inciting tensions over the Jerusalem shrine. The government has no intention of creating any kind of change in the status quo at Temple Mount, he said. Israeli and Palestinian peace talks have been frozen for over a year, after the latest round of U.S.-mediated negotiations collapsed in April, 2014.

Therefore class 5 articles have to be classified using unrelated to commodities labels. In doing so most used words lists can help in context extraction and labeling.

Conclusion

As evidenced by the results, approximately ¼ of the articles deal with agriculture. However it has to be noted that even agriculture unrelated articles may have an effect on prices.

The fact that wheat is the most covered commodity in agriculture-related articles is confirmed by this analysis, but it has to be underlined that many class 0 articles may contain sentiment related to other commodities, such as maize. Under this point of view a multi-labeling approach is suggested.

Wheat and Maize classes (class 0 and class 1) show some degree of format noisiness and articles may talk about different topics, as export quantities/values or future market status.

Next research topics:

- Given articles noisiness, it could be a good idea to format the text in a plainer format by removing, for example, the parts of the articles related to webpage formatting (“ …1px;\r background-color: #FFFFFF… ”).
- Word lists can be refined.
- A sharp error measurement method has to be defined.
- A more structured TF/IDF approach has to be applied.

- Usage as additional labels/features of the most common words (Word1, Word2, Word3, Word4 and Word5). Such words can also be used not only for commodity labeling, but also in other classification tasks.
- Once satisfying results are obtained on a sample, the same sample can be used as a training set for a classifier.